Environmental Health Perspectives
Vol. 63, pp. 217-224, 1985

Use of Historical Controls for Animal Experiments

by Takashi Yanagawa* and David G. Hoel†

Statistical methods for the use of historical control data in testing for a trend in proportions in carcinogenicity rodent bioassays are reviewed. Asymptotic properties of the Hoel-Yanagawa exact conditional tests are developed and compared with the Tarone test. It is indicated that the Hoel-Yanagawa test is more powerful than the Tarone test. These tests depend on the beta-binomial parameters which are estimated from historical data. The goodness of fit of beta-binomial distributions to historical data is illustrated by application to the historical control database in the National Toxicology Program. Finally, sensitivities of the exact conditional test to the historical information is discussed and a conservative use of the test is considered.

Introduction

To begin, we consider Table 1, which summarizes the data from an experiment involving $r + 1$ groups of animals. One group serves as a control group and the remaining r groups are administered a test compound at increasing dose levels, $d_1 < d_2 < \ldots < d_r$. The control group is associated with $i = 0$ so that $d_0 = 0$. Let n_i denote the number of animals in the i-th group. We assume for $i = 0, 1, \ldots, r$ that at experimental dose d_i there are x_i animals with tumors observed which are binomially distributed with parameters p_i and n_i. We define $p = p_0$.

To test an increase in the proportions $p_i = x_i/n_i$ with increasing dose level, Cochran (*1*) and Armitage (*2*) suggested the test statistic

$$X^2 = \frac{\left(\sum_{i=1}^{r} x_i d_i - \hat{p}\sum_{i=1}^{r} n_i d_i\right)^2}{\left\{\hat{p}\hat{q}\left[\sum_{i=1}^{r} n_i {d_i}^2 - \left(\sum_{i=1}^{r} n_i d_i\right)^2 / n\right]\right\}}$$

where $\hat{p} = x/n$ and $\hat{q} = 1 - p$. This statistic is distributed asymptotically as a chi-squared random variable with one degree of freedom if there are no differences in the probability p_i of developing a tumor among the $r+1$ groups. Cox (*3*) showed that this statistic gives the uniformly most powerful unbiased test against logistic alternatives and Tarone and Gart (*4*) showed that this statistic is asymptotically locally optimum against any alternative which can be expressed as a smooth increasing function of dose.

Table 1. Summary data from an animal carcinogenesis bioassay.

	Dose level					
	0	d_1	d_2	...	d_r	Total
Animals with tumor	x_o	x_1	x_2	...	x_r	x
Animals without tumor	$n_o - x_o$	$n_1 - x_1$	$n_2 - x_2$	...	$n_r - x_r$	$n - x$
Sample size	n_o	n_1	n_2	...	n_r	n

In most carcinogenicity rodent bioassays, we are usually dealing with three experimental groups of animals which consist of a control group, a low dose and a high dose group each with 50 animals. The probability of an animal with a specific type of tumor in the control group ranges from less than 1% to 20% depending upon the type of tumor.

When the Cochran-Armitage test is applied to these bioassays, two problems arise: the problem of false positives (Type I error) and that of false negatives (Type II error). For the first problem, Portier and Hoel (*5*) showed that when the Cochran-Armitage test is used the false positives can be considerable, depending mark-

*Department of Mathematics, Kyushu University 33, Fukuoka 812, Japan.

†Radiation Effect Research Foundation, Hiroshima, Japan, and Biometry and Risk Assessment Program, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709.

Table 2. False negative rate (Type II error) of the Cochran-Armitage test at 5% nominal level, $r = 2$, $n_0 = n_1 = n_2 = 50$, and $d_1 = 1$, $d_2 = 2$.

	1	2	3	4	5	6	7	8
p	0.01	0.01	0.01	0.05	0.05	0.05	0.10	0.10
p_1	0.015	0.022	0.049	0.073	0.106	0.212	0.143	0.201
p_2	0.022	0.049	0.360	0.106	0.212	0.580	0.201	0.363
False negative rate[a]	0.89	0.80	0.51	0.77	0.44	0.03	0.66	0.21

[a] Computed from the asymptotic power of the Cochran-Armitage test.

edly on bioassay design. This problem can be overcome by the use of the exact trend test which extends Fisher's exact test for a 2×2 table. Yanagawa, Hoel, and Brooks (*6*) show that the computing time for the exact trend test is fairly short. However, whether we use the Cochran-Armitage test or the exact trend test, the second problem, that is the problem of the false negatives, still remains. Table 2 shows approximate false negative rates of the trend test at nominal 5% significance level for several values of (p,p_1,p_2). The values in the table are approximated by means of the asymptotic power of the Cochran-Armitage test. The table shows that the false negative rates are fairly large. For example, when $p = 0.01$, $p_1 = 0.05$, and $p_2 = 0.36$, the table reveals that approximately one-half the time a carcinogen is tested and found to be noncarcinogenic by the trend tests.

Clearly, the need is to increase the power of the trend tests so as to decrease the false negative rate. Now, the National Cancer Institute (NCI) and the National Toxicology Program (NTP), USA, have generated nearly 300 Technical Reports summarizing the results of carcinogenicity rodent bioassays for a wide variety of chemicals. For each of these studies detailed information of neoplastic and nonneoplastic lesions for individual animals have been computerized and stored on the Carcinogenesis Bioassay Data System. This replication of experiments leads to a general knowledge by toxicologists of what outcomes are typically observed in an experimental control group.

The purpose of this paper is to review statistical methods for the use of historical control data in testing for a trend in proportions in carcinogenicity rodent bioassays. Asymptotic properties of the Hoel-Yanagawa (*7*) exact conditional tests are developed. The asymptotic conditional test is compared with the Tarone (*8*) test and shown to have higher power than the latter. The goodness of fit of beta-binomial distributions to historical data which is the basic assumption for the Hoel-Yanagawa and Tarone tests is examined by using the historical control database established by the NTP. Finally, sensitivities of the exact conditional test to the historical information is considered and a conservative use of the test is discussed.

Literature Review

Several authors have developed methods for incorporating historical information into statistical tests of hypothesis. Tarone (*8*) assumes a logistic dose–response model with a beta prior distribution for the probability of an animal with a tumor in the control group. Using likelihood methods an asymptotic test is developed for determining the existence of a positive dose-response. The test functionally is a modification of the Cochran-Armitage test. Dempster, Syelwyn, and Weeks (*9*) assume that the logic of the historical rates are normally distributed and apply Bayesian methods to obtain a p-value of the posterior probability of a positive dose-response. The authors indicate that their approach is asymptotically equivalent to the large sample test of Tarone. Hoel (*10*) proposes a conditional two-sample test. His idea is more thoroughly discussed in Hoel and Yanagawa (*7*). The tests developed are exact tests rather than the asymptotic procedures of the previous authors. They assume, as did Tarone (*8*), that the historical rates are distributed as beta-binomial and construct tests conditional on the number of outcomes in the control group.

In all of this work, the parameters of the beta-binomial distribution which must be estimated from the historical control data are assumed to be known. Considering the conditional test for logistic response by Hoel and Yanagawa (*7*), Yanagawa, Hoel, and Brooks (*6*) discuss the sensitivity of the test to this source of variability and develop a conservative use of the test. Haseman, Huff, and Boorman (*11*) have reviewed the data stored in the Carcinogenesis Bioassay Data System and discuss those issues which must be adequately addressed before historical control data can be used in a formal testing framework.

Formulation

We assume for $i = 0, 1, \ldots, r$ that at experimental dose d_i there are X_i animals with observed tumors which are assumed to be binomially distributed with parameters p_i and n_i. We formulate the problem of testing for a trend in proportions following Tarone and Gart (*4*) by assuming that:

$$p_i = H(a + \xi d_i)$$

where H is a twice differentiable and monotone increasing function over $[0,\infty]$. The statistical test of hypothesis of an increasing trend in proportions is given by

$$H_0: \xi=0 \quad \text{vs.} \quad H_1: \xi>0$$

Following the development of Tarone (*8*) and Hoel (*10*), we assume that p_0 (denoted by p) is a random variable following a beta distribution

$$g(p) = \Gamma(\alpha + \beta)p^{\alpha-1}q^{\beta-1}/[\Gamma(\alpha)\Gamma(\beta)]$$

with

$$q = 1 - p$$

with α and β known. We defined $p = H(a)$ so that a is distributed as

$$f(a) = \Gamma(\alpha+\beta)H(a)^{\alpha-1}[1-H(a)]^{\beta-1}H'(a)/[\Gamma(\alpha)\Gamma(\beta)]$$

Since $X_0, X_1, \ldots, X_r$ are independent conditioned on p, the joint distribution of a and $X = (X_0, X_1, \ldots, X_r)$ is given by

$$f_\xi(a,x) = \prod_{i=0}^{r} \binom{n_i}{x_i} H(a + \xi d_i)^{x_i}[1 - H(a + \xi d_i)]^{n_i - x_i} f(a)$$

Thus the marginal distribution of X is

$$f_\xi(x) = \int_{-\infty}^{\infty} f_\xi(a,x)da$$

In particular, the marginal distribution of X_0 is

$$f_\xi(x_0) = \binom{n_0}{x_0} \Gamma(\alpha + \beta)\Gamma(x_0 + \alpha)\Gamma(n_0 + \beta - x_0)/ \{\Gamma(\alpha)\Gamma(\beta)\Gamma(n_0 + \alpha + \beta)\}$$

which is independent of ξ.

Unconditional Tests

The locally most powerful test (*12*) for H_0: $\xi = 0$ vs. H_1: $\xi > 0$ is given by

$$T = \frac{d \log f_\xi(x)}{d\xi}\bigg|_{\xi=0} = \frac{f_\xi'(x)}{f_0(x)}\bigg|_{\xi=0}$$

After some simple calculation we find that

$$T = \frac{\Gamma(n + \alpha + \beta)}{\Gamma(x + \alpha)\Gamma(n + \beta + x)} \int_{-\infty}^{\infty} \left[\sum_{i=1}^{r} x_i d_i - \sum_{i=1}^{r} n_i d_i H(a)\right] H(a)^{x+\alpha-2}[1 - H(a)]^{n+\beta-x-1}[H'(a)]^2 da \tag{1}$$

which depends on the response function H. Two cases are of particular interest:

(1) Logistic response: Set $H(a) = e^a/(1+e^a)$, then from Eq. (1) we have

$$T_t(\alpha,\beta) = \sum_{i=1}^{r} x_i d_i - \frac{x + \alpha}{n + \alpha + \beta} \sum_{i=1}^{r} n_i d_i \tag{2}$$

(2) Exponential response: Set $H(a) = 1 - e^{-a}$, then

$$T_h(\alpha,\beta) = \frac{n + \alpha + \beta - 1}{x + \alpha - 1} \sum_{i=1}^{r} x_i d_i - \sum_{i=1}^{r} n_i d_i = \frac{n + \alpha + \beta - 1}{x + \alpha - 1} T_t(\alpha - 1,\beta)$$

Suppose that the response function is a distribution function which is third-order differentiable, then applying the formula by Hald (*13*), we may show that

$$T = \frac{(n + \alpha + \beta)^2}{(x + \alpha)(n + \beta - x)} w\left(\frac{x + \alpha}{n + \alpha + \beta}\right) T_t + o_p(\sqrt{n})$$

where $w(h) = H'[H^{-1}(h)]$ and T_t is the statistic given in Eq. (2). Thus when T and T_t are appropriately normalized, they have the same asymptotic distribution as $n \to \infty$; that is, T is asymptotically free of the shape of the response function and is equivalent to T_t.

In order to obtain an asymptotic test based on T_t we observe the following results which are straightforward calculations. Under H_0: $\xi = 0$

$$E(X_i) = n_i\alpha/(\alpha + \beta) = n_i\theta$$
$$V(X_i) = n_i\theta(1 - \theta)/(\alpha + \beta + 1)$$
$$\text{Cov}(X_i, X_j) = n_i n_j\theta(1 - \theta)/(\alpha + \beta + 1)$$

and it thus follows that

$$E(T_t) = 0$$

$$V(T_t) = \frac{\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)} \left\{\sum_{i=1}^{r} n_i {d_i}^2 - \frac{1}{n + \alpha + \beta}\left(\sum_{i=1}^{r} n_i d_i\right)^2\right\}$$

When conditioned on p, the mean and variance of T_t under H_0 are

$$E(T_t \mid p) = \frac{p(\alpha + \beta) - \alpha}{n + \alpha + \beta} \sum_{i=1}^{r} n_i d_i$$

$$V(T_t \mid p) = pq\left\{\sum_{i=1}^{r} n_i {d_i}^2 - \frac{n + 2(\alpha + \beta)}{(n + \alpha + \beta)^2}\left(\sum_{i=1}^{r} n_i d_i\right)^2\right\}$$

The test statistic T_t could be standardized as

$$S_c = T_t/V(T_t \mid p)^{1/2}$$

or as

$$S = T_t/V(T_t)^{1/2}$$

with p and q replaced by their estimates $\hat{p} = (X/n)$ and $\hat{q} = 1 - \hat{p}$ or as Tarone did with

$$S_t = T_T/\left\{\frac{(X + \alpha)(n + \beta - X)}{(n + \alpha + \beta)^2}\left[\sum_{i=1}^{r} n_i d_i^2 - \frac{1}{(n + \alpha + \beta)}\left(\sum_{i=1}^{r} n_i d_i\right)^2\right]\right\}^{1/2}$$

Of these standardizations, the first statistic S simply uses the unconditional variance of the statistic T while the second uses the estimated conditional variance. The Tarone standardization results from treating the random variable a as a parameter in the likelihood function and using the score test of $\xi = 0$.

In considering the asymptotic distribution of the test statistics S, S_c, S_t as $n \to \infty$ we assume that $\lambda_i = n_i/n$ is kept constant $(0 < \lambda_i < 1)$ for each i. The asymptotic distributions are summarized as follows and the proofs can be found or obtained following the arguments given in Hoel and Yanagawa (7):

TARONE'S STANDARDIZATION. For either $\alpha + \beta$ or $\theta = \alpha/(\alpha + \beta)$ fixed and $\Phi(\cdot)$ the normal distribution, then under H_0

$$\lim_{n\to\infty} \text{pr}\{S_t < x\} = \Phi(x)$$

UNCONDITIONAL STANDARDIZATION. For $\alpha + \beta$ fixed then

$$\lim_{n\to\infty} pr\{S < x\} = E\{\Phi[x\sqrt{\theta\beta/pq(\alpha + \beta + 1)}]\}$$

For $\alpha + \beta \to \infty$, then

$$\lim_{n\to\infty} \text{pr}\{S < x\} = \Phi(x)$$

STANDARDIZATION BY CONDITIONAL VARIANCE

$$\lim_{n\to\infty} \text{pr}\{S_c < x\} = \Phi(x/\sqrt{k})$$

where

$$k = \frac{\Sigma\lambda_i d_i^2 - \rho(\Sigma\lambda_i d_i)^2}{[\Sigma\lambda_i d_i^2 - \rho(2 - \rho)(\Sigma\lambda_i d_i)^2]}$$

where $\rho_n = n/(n + \alpha + \beta) \to \rho$ $(0 < \rho < 1)$ and $\lambda_i = n_i/n$ is fixed.

The above results show that Tarone's standardization is the best among the three, although the standardization is not easy to justify. It is shown in Hoel and Yanagawa (7) that when θ is small, n must be quite large for the normality of the asymptotic tests to be a reasonable approximation.

Exact Conditional Test

Since $f_0(x)$ is independent of ξ, we have that X_0 is an ancillary statistic. Fisher (14) suggested that for purposes of inference on should consider the family of conditional distributions given the observed value of the ancillary statistic in the sample. Denote by $f_{\xi}(x|x_0)$ the conditional probability density function of X given $X_0 = X_0$. The conditional locally most powerful test for H_0: $\xi = 0$ vs. H_1: $\xi > 0$ is given by

$$T = [d \log f_\xi(x \mid x_0)/d\xi]_{\xi=0}$$

and it is easy to show that T is given by Eq. (1).

In general, let t_0 be the observed value of T; then the exact p-value of the conditional test is given by

$$p\text{-value} = \sum{}' \prod_{i=1}^{r} \binom{n_i}{x_i} \frac{\Gamma(x + \alpha)\Gamma(n + \beta - x)\Gamma(n_0 + \alpha + \beta)}{\Gamma(x_0 + \alpha)\Gamma(n_0 + \beta - x_0)\Gamma(n + \alpha + \beta)}$$

where the summation Σ' extends over all $(x_1, x_2, \ldots, x_r)$, which satisfy $T < t_0$ for given $X_0 = x_0$. For the NTP data with $r = 2$, $n_0 = n_1 = n_2 = 50$ and p ranging from 1% to 20%, computations of the p-value by computer is very quick.

Asymptotic Properties of the Exact Conditional Tests

Hoel and Yanagawa (7) standardized the statistic T_t by means of the conditional expectation and variance given X_0 and obtained the following statistic S:

$$S = T_t/\{\mu(1 - \mu)[\Sigma\, n_i d_i^2 - \frac{1}{n + \alpha + \beta}(\Sigma\, n_i d_i)^2]\}^{1/2}$$

where

$$\mu = (x_0 + \alpha)/(n_0 + \alpha + \beta).$$

Following the development in Hoel and Yanagawa (7) one may show that: (1) under the null hypothesis H_0, S has limiting normal distribution with mean zero and variance one as $n \to \infty$. Thus the asymptotic conditional size α test for increasing trend in proportion is given by rejecting the hypothesis H_0 if $S \geq z_\alpha$, where z_α is the upper α% point of the standard normal distribution.

(2) For the sequence of alternative hypotheses H_{1n}: $\xi_n = \delta/\sqrt{n}$ the unconditional asymptotic power of the conditional test S is the same as for S_t which is given in Theorem 2 of Hoel and Yanagawa (*7*). (3) The Pitman asymptotic relative efficiency of the conditional test with respect to the Cochran-Armitage test is

$$\text{ARE}(S \mid S_{C-A}) = [B(\rho)/B(1)]^2$$

This efficiency formula can be used for assessing the saving in sample size by incorporating historical controls. Suppose that historical control data are incorporated with the current experiment which uses n total animals. Then the formula implies that approximately $n' = n[B(\rho)/B(1)]^2$ animals are needed by the analysis of the current experiment alone to achieve the same statistical power as the incorporated analysis. For example, the analysis of incorporating historical controls of $\alpha+\beta = 400$ with the current experiment using $d_1 = 1$, $d_2 = 2$, and $n_0 = n_1 = n_2 = 50$ animals corresponds to the analysis of $n_0 = n_1 = n_2 = 105$ animals of current experiment. In terms of the false negatives of statistical tests, the use of this historical information decreases the false negative rate from 0.51 to 0.24 when $p = 0.01$, $p_1 = 0.049$, and $p_2 = 0.360$ (see Table 2). This of course assumes that $\alpha+\beta \to \infty$ as $n \to \infty$ and that $p = \theta$ in the limit. If this is not the case, the above finding is not true. For example, in comparing the exact p-values of the exact tests, Yanagawa, Hoel, and Brooks (*6*) show that when θ is large ($\theta = 0.2$), $\alpha+\beta$ is small ($\alpha+\beta = 15$) and x_0/n_0 is much smaller than θ; then the p-value of the Hoel-Yanagawa conditional test for $(x_0,x_1,x_2) = (2,2,9)$ is 0.048; whereas the corresponding p-value of the exact trend test which does not incorporate historical control data is 0.0096. Generally, Hoel and Yanagawa (*7*) find that the Cochran-Armitage test gives much higher p-values especially when x_0 is larger than expected and smaller p-values when x_0 is smaller than expected.

Comparisons with the Tarone Test

Suppose that the sample size n and p are moderately large and that the distributions of both test statistics S and S_t are approximated well by their asymptotic distributions. It would be reasonable to expect under the alternative hypothesis of a positive dose-response that the observed sample point $(x_0, x_1, \ldots x_r)$ falls in the region R defined by

$$R = \{x_0,x_1,\ldots,x_r) : x_0/n_0) \leqslant (x_i/n_i),\ i = 1,2,\ldots,r\}$$

Furthermore assume that $x_i/n_i < 1/2$, $i = 0, 1, \ldots, r$, which is the case in many animal carcinogenicity experiments. Then it may be shown for $(x_0,x_1, \ldots, x_r) \in R$ that S is larger than the square root of Tarone's test statistic. This indicates for a moderate sample size that the asymptotic conditional test would have higher power than the Tarone test.

Generally, as stated above for animal experiments where n is small, the asymptotic approximation of the trend tests is not good, especially when θ is very small and $\alpha+\beta$ is large. This is the situation where a good gain in power by incorporating historical control data is anticipated. Therefore, the exact conditional test is suggested rather than asymptotic test.

Finally, we note one weak point of the Tarone test, as well as the test by the other authors. Suppose that $n_0 = n_1 = n_2 = 50$, $\theta = 0.01$ and $(\alpha, \beta) = (3.95, 3.91)$, and that $(x_0,x_1,x_2) = (3,3,3)$ is observed, then the p-value of the Tarone test is 0.007. Thus a strong evidence of positive dose-response is shown. This is because we have $\text{pr}[x_0 \geqslant 3] = 0.02$. This illustrates the necessity of dealing with exceptional values of x_0 which happen sometimes by the reasons discussed in the next section. Since the existence of sound historical control database has been presumed for our statistical procedures, one should not attempt to incorporate the historical data when the exceptional value of x_0 is observed. We encourage the use of the ancillary information, i.e. x_0, in the conditional procedure to check the quality of current experiment.

Historical Control Database

Problems encountered in the historical control data are discussed by Haseman, Huff, and Boorman (*11*). Examining the NCI/NTP historical data carefully, these authors find that different terminologies are often used to describe the same tumor even for studies at the same laboratory carried out at approximately the same time. Also the use of different sets of criteria for diagnosing a lesion is revealed. Discussing the criteria that will aid in determining whether a particular study should be included in the database, Haseman, Huff, and Boorman (*11*) state "Certainly species, strain, sex, study duration, pathology protocols, nomenclature conventions, quality assurance and review procedures should be the same for each study in a particular control database. Ideally, diets, changing regiments, and various environmental parameters should also be comparable. Different types of control groups (e.g., untreated, corn oil gavage) should be dealt with separately. Other potential sources of variability (calendar year, laboratory, pathologist, supplier) should also be investigated, identified and controlled." The current database thus established in the NTP contains information beginning with those studies reported in Technical Report 193, 1981 through those studies whose pathology diagnoses were finalized in Carcinogenesis Bioassay Data System as of March, 1983. Most control groups have 50 animals/species/sex and all are from studies of two years duration. About 50 control groups/species/sex are contained in the database.

We fitted beta-binomial distribution to the data for each tumor type in the database. Table 3 shows for selected tumor sites in the Fisher 344 rat the estimates of the beta-binomial parameters α and β, $\alpha+\beta$, and $\theta = \alpha/(\alpha+\beta)$, and their standard deviations. These es-

Table 3. Carcinogenesis data system: maximum likelihood estimates of beta-binomial parameters in F344 rats.[a]

	Male				Female			
	0	α	β	α + β	0	α	β	α + β
Respiratory system	0.023	14.7	626.0	641.0	0.011	2.8	242.3	245.1
Lung nos.	(0.003)[b]	(50.9)	(2170.5)	(2221.3)	(0.002)	(3.2)	(284.1)	(287.3)
Alveolar/bronchiolar adenoma								
Alveolar/bronchiolar carcinoma								
Total incidence/animals	53/2305				27/2354			
Hematopoietic system	0.302	9.0	20.9	29.9	0.189	15.7	67.5	83.3
All lymphomas	(0.015)	(3.0)	(7.1)	(10.1)	(0.010)	(8.8)	(38.1)	(45.9)
All leukemias								
Total incidence/animals	699/2320				448/2370			
Circulatory system	0.007	*			0.002	*		
Hemangioma	(0.002)				(0.001)			
Hemangiosarcoma								
Angioma								
Angiosarcoma								
Total incidence/animals	16/2320				5/2320			
Digestive system	0.042	2.5	58.0	60.5	0.031	1.8	55.0	56.8
Liver nos.	(0.006)	(1.3)	(29.9)	(3.1)	(0.005)	(1.0)	(30.8)	(31.7)
Hepatocellular adenoma								
Neoplastic nodule								
Hepatocellular carcinoma								
Total incidence/animals	96/2306				74/2356			
Endocrine system	0.089	8.1	83.5	91.6	0.084	10.1	110.7	120.8
Thyroid/thyroid follicle	(0.007)	(4.9)	(51.0)	(55.8)	(0.007)	(7.6)	(83.7)	(91.2)
C-cell adenoma								
C-cell carcinoma								
Total incidence/animals	196/2230				189/2265			
Endocrine system	0.180	6.6	25.4	31.0	0.039	*		
Adrenal: nos./capsule/ cortex/medulla	(0.013)	(1.9)	(8.7)	(10.6)	(0.004)			
Pheochromocytoma								
Pheochrocytoma, malignant								
Total incidence/animals	409/2280				92/2338			
Reproductive system	0.022	*			0.238	6.3	20.1	26.4
Mammary gland	(0.003)				(0.014)	(2.0)	(6.6)	(8.6)
Adenoma, nos.								
Papillary adenoma								
Cystadenoma, nos.								
Papillary cystadenoma, nos.								
Intraductal papilloma								
Acinar-cell adenoma								
Fibroma								
Fibroadenoma								
Total incidence/animals	52/2320				564/2370			
Urinary system	0.003	*			0.002	*		
Kidney nos.	(0.001)[b]				(0.001)			
Adenoma, Nos.								
Tubular-cell adenoma								
Papillary cystadenoma, nos.								
Adenocarcinoma, nos.								
Tubular-cell adenocarcinoma								
Total incidence/animals	6/2307				4/2359			
Endocrine system	0.058	15.1	245.3	260.4	0.010	0.8	74.1	74.9
Pancreatic islet	(0.005)	(21.5)	(349.8)	(371.3)	(0.003)	(0.6)	(57.2)	(57.7)
Islet-cell adenoma								
Islet-cell carcinoma								
Total incidence/animals	129/2226				24/2303			

Table 3. (contd).

	Male				Female			
	0	α	β	α + β	0	α	β	α + β
Endocrine system	0.017	3.0	167.1	170.1	0.008	*		
Thyroid/thyroid follicle	(0.003)	(3.1)	(178.4)	(181.5)	(0.002)			
Follicular-cell adenoma								
Cystadenoma, nos.								
Papillary cystadenoma, nos.								
Papillary carcinoma								
Adenocarcinoma, nos.								
Papillary adenocarcinoma								
Follicular-cell carcinoma								
Papillary cystadenocarcinoma, nos.								
Total incidence/animals	39/2230				19/2265			

[a] The asterisk (*) shows that a beta-binomial distribution did not fit to the data.
[b] Standard deviation.

timates are obtained by the method of maximum likelihood. The asterisk (*) in the table represents tumor sites whose data did not fit well to a beta-binomial distribution. It is estimated that there are a little more than 1/3 of such in the database. In most of these, data variations between experiments are rather smaller than that of a binomial distribution. Methods for incorporating these historical control data are not yet developed. Figures are given in Yanagawa, Hoel, and Brooks (*6*) to show visually the goodness-of-fit of beta-binomial distributions to several selected tumor sites.

Sensitivity

The exact conditional tests developed in the preceding sections depend on beta-binomial parameters, α and β, which are estimated from historical control data. As seen in Table 3, the standard deviations of the estimated α and β are fairly large. There is the need to consider the effect of this source of variability.

Yanagawa, Hoel, and Brooks (*6*) studied its effect on the *p*-value of the conditional test for a logistic response. They show that it changes only slightly with a small change in α + β; that when α + β is small the *p*-value is not sensitive to a change in θ; whereas when α + β is large, slight changes in θ produce substantial changes in the *p*-value, and in particular, when the difference of θ and x_0/n_0 is large.

Developing methods for constructing the 95% confidence intervals of θ and α + β, Yanagawa, Hoel, and Brooks (*6*) considered the maximum and minimum of *p*-values over the space made by the cartesian product of these confidence intervals (see the shaded area in Figure 1). They found numerically that these maxima and minima seem to be attained always at the four corner points,

Table 4. The *p*-value at the points *A*, *B*, *C*, *D*, and 0.

Tumor	(X_0,X_1,X_2)			*p*-value				
				0	*A*	*B*	*C*	*D*
Thyroid/thyroid follicle[a]	0	0	4	0.016	0.003	0.009	0.004	0.041
(α,β) = (2.98,167.46)	0	1	4	0.010	0.003	0.008	0.002	0.026
(θ, α + β)	0	2	3	0.023	0.006	0.016	0.005	0.056
	1	0	4	0.033	0.040	0.057	0.004	0.045
	1	1	4	0.022	0.036	0.052	0.002	0.029
0: (0.017 170.4)	1	2	4	0.015	0.031	0.044	0.001	0.017
A: (0.11, 21.7)	2	0	5	0.020	0.056	0.072	0.001	0.013
B: (0.024, 21.7)	2	2	4	0.028	0.027	0.064	0.013	0.040
C: (0.011, 1445.1)	2	3	4	0.018	0.020	0.050	0.008	0.026
D: (0.024, 1334.1)	3	4	5	0.008	0.089	0.108	0.002	0.009
Hematopoietic system[b]	4	4	15	0.029	0.013	0.040	0.017	0.055
(α,β) = (15.8,67.62)	4	8	14	0.029	0.014	0.042	0.017	0.053
(θ, α +β)	4	12	14	0.015	0.005	0.011	0.010	0.055
	8	4	17	0.034	0.018	0.049	0.020	0.058
0: (0.189, 83.4)	8	8	16	0.034	0.020	0.050	0.020	0.056
A: (0.169, 27.7)	8	12	15	0.033	0.030	0.048	0.011	0.056
B: (0.209, 27.7)	12	12	18	0.022	0.039	0.059	0.003	0.018
C: (0.169, 251.1)	12	12	17	0.021	0.048	0.072	0.010	0.032
D: (0.209, 251.1)	16	16	21	0.008	0.036	0.054	0.000	0.002

[a] Includes follicular-cell adenoma; cystadenoma, nos; papillary cystadenoma, nos; papillary carcinoma; adenocarcinoma, nos; papillary adenocarcinoma; follicular-cell carcinoma; papillary cystadenocarcinoma, nos.
[b] Includes all lymphomas and leukemias.

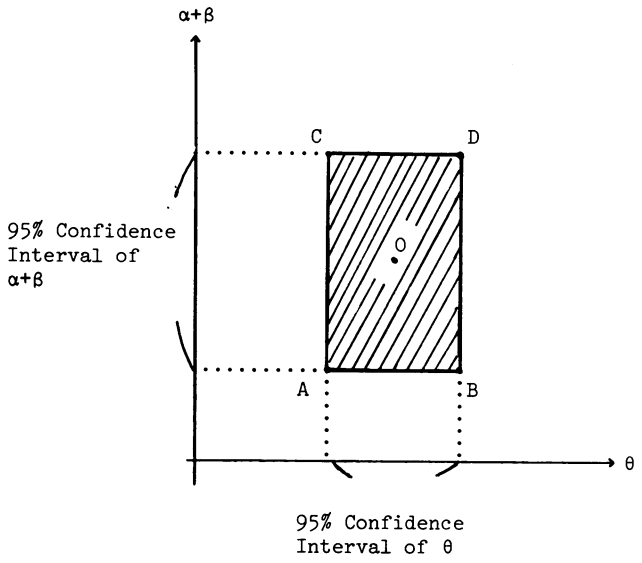

FIGURE 1. Area made by the 95% confidence intervals of θ and α + β.

i.e., A, B, C, and D in Figure 1. Table 4 shows p-values at A, B, C, and D, and at the point 0 of estimated α and β for several configurations of (x_0,x_1,x_2) for tumors of the thyroid and tumors of the hematopoietic system using as usual $n_0 = n_1 = n_2 = 50$.

Conservative Use of the Conditional Test

The inspection of Table 4 leads to the following conservative rule for incorporating historical control data by the exact conditional test for testing positive dose-response:

(R1) Compute p-values at the five points A, B, C, D, and O.

(R2) Do not attempt to draw any inference when the maximum p-value of these five points exceeds the nominal level, e.g., 0.01 or 0.05.

This rule is very conservative, but it still works well in practice, especially for tumors with small spontaneous background rates. This is shown by comparing the maximum p-value with the p-value of the exact trend test, i.e., extended version of Fisher's exact test which does not incorporate historical data. For example, when $(\hat{\alpha},\hat{\beta}) = (3.95,391)$ and $(x_0,x_1,x_2) = (1,2,3)$, then the maximum p-value is 0.020; whereas the p-value of the exact test is 0.226. The rule also works for many configurations of (x_0,x_1,x_2) even when θ is large ($\theta = 0.2$); for example, when $(\hat{\alpha},\hat{\beta}) = (3,12)$ and $(x_0,x_1,x_2) = (21,25,29)$, then the maximum p-value is 0.018 and the p-value of the exact test is 0.067. Note that the computing time required to obtain the p-values at the five points is rather short: for example, when $(\hat{\alpha},\hat{\beta}) = (3.95,391)$ a VAX 780 took less than 40 sec to compute the p-values for $(x_0,x_1,x_2) = (1,2,3)$.

REFERENCES

1. Cochran, W. G. Some methods for strengthening the common χ^2 tests. Biometrics 10: 417–451 (1954).
2. Armitage, P. Tests for linear trends in proportions and frequencies. Biometrics 11: 375–386 (1955).
3. Cox, D. R. The regression analysis of binary sequences (with discussion). J. Roy. Statist. Soc. B20: 215–242 (1958).
4. Tarone, R. E., and Gart, J. On the robustness of combined tests for trends in proportions. J. Am. Statist. Assoc. 75: 110–116 (1980).
5. Portier, C., and Hoel, D. G. Type I error of trend tests in proportions and the design of cancer screens. Submitted .
6. Yanagawa, T., Hoel, D. G., and Brooks. G. T. A conservative use of historical data for a trend test in proportions. Submitted.
7. Hoel, D. G., and Yanagawa, T. Incorporating historical controls in testing for a trend in proportions. Submitted.
8. Tarone, R. E. The use of historical control information in testing for a trend in proportion. Biometrics 38: 215–220 (1982).
9. Dempster, A. P., Selwyn, M. R., and Weeks, B. J. Combining historical and randomized controls for assessing trends in proportions. J. Am. Statist. Assoc. 78: 221–227 (1983).
10. Hoel, D. G. Conditional two-sample tests with historical controls. In: Contributions to Statistics: Essays in Honour of Norman L. Johnson (P. K. Sen, Ed.), North Holland, Amsterdam, 1983, pp. 229–236.
11. Haseman, J. K., Huff, J., and Boorman, G. A. Use of historical control data in carcinogenicity studies in rodent. Submitted .
12. Rao, C. R. Linear Statistical Inference and Its Application. John Wiley & Sons, New York, 1973.
13. Hald, A. The mixed binomial distribution and posterior distribution of p for a continuous prior distribution. J. Roy. Statist. Soc. B30: 359-367 (1968).
14. Fisher, R. A. Statistical Methods and Scientific Inference. Oliver and Boyd, London, 1956.